

Organización de Datos 75.06. Primer Cuatrimestre de 2018. Examen parcial, primera oportunidad:

Importante: Antes de empezar complete nombre y padrón en el recuadro. Lea bien todo el enunciado antes de empezar. Para aprobar se requiere un mínimo de 60 puntos (60 puntos = 4) con al menos 20 puntos entre los ejercicios 1 y 2. Este enunciado debe ser entregado junto con el parcial si quiere una copia del mismo puede bajarla del grupo de la materia. En el ejercicio 3 elija 2 de los 4 ejercicios y resuelva únicamente 2 ejercicios. Si tiene dudas o consulta levante la mano, está prohibido hablar desde el lugar, fumar o cualquier actividad que pueda molestar a los demás. El criterio de corrección de este examen está disponible en forma pública en el grupo de la materia.

"You either die a hero or live long enough to see yourself become the villain." Harvey Dent, *The Dark Knight Rises*

#	1	2	3.1/1	3.2/1	4	5	6	7	Entrega Hojas:
Corrección									Total:
Puntos	/15	/15	/15	/15	/15	/15	/15	/15	_____ /100

Nombre:
Padrón:
Corregido por:

<p>1) Nintendo of America (EEUU) tiene información de ventas de videojuegos físicos mensuales totalizadas en EEUU las cuales se realizan en cadenas de tiendas de videojuegos en el siguiente RDD: (<i>id_videojuego</i>, <i>id_tienda</i>, mes, año, <i>total_ventas_mensuales</i>).</p> <p>Por otro lado tenemos un RDD con información de las tiendas y de su ubicación (<i>id_tienda</i>, dirección, latitud, longitud, código_postal, estado).</p> <p>Con esta información escribir un programa en pySpark para obtener la tienda que realizó menor cantidad de ventas en el estado de "Georgia" en todo el año 2017. (***) (15 pts)</p>	<p>2) (****) (15 pts) El GCPD (Gotham City Police Dept) recolecta la información de casos policiales que acontecen en Ciudad Gótica. Esta información se encuentra guardada en un dataframe con el siguiente formato: (<i>fecha</i>, <i>id_caso</i>, <i>descripción</i>, <i>estado_caso</i>, <i>categoria</i>, <i>latitud</i>, <i>longitud</i>).</p> <p>Los posibles estados que puede tener un caso son 1: caso abierto, 2: caso resuelto, 3: cerrado sin resolución. Las fechas se encuentran en el formato YYYY-MM-DD.</p> <p>Por otro lado el comisionado Gordon guarda un registro detallado sobre en cuáles casos fue activada la batiseñal para pedir ayuda del vigilante, Batman. Esta información se encuentra en un Dataframe con el siguiente formato (<i>id_caso</i>, <i>respuesta</i>), siendo campo respuesta si la señal tuvo una respuesta positiva (1) o negativa (0) de parte de él.</p> <p>El sector encargado de las estadísticas oficiales del GCPD quiere con esta información analizar las siguientes situaciones:</p> <ul style="list-style-type: none"> - Tasa de resolución de casos de la fuerza policial por categoría de caso (considerando aquellos casos en los que no participó Batman). - Tasa de resolución de casos con la ayuda de Batman (considerando que aquellos casos en los que fue llamado con la batiseñal, participó en la resolución). - Indicar el mes del año pasado en el que Batman tuvo mayor participación en la investigación de casos.
---	---

3) Resolver 2 (dos) y solo 2 de los siguientes ejercicios a elección (si resuelve más de 2 el ejercicio vale 0 puntos, sin excepciones). En cada caso indicar V o F justificando adecuadamente sus respuestas. Si no justifica vale 0 puntos sin excepciones.

a) Al reducir dimensiones utilizando la descomposición por valores singulares lo que buscamos es quedarnos con los menores valores singulares que serán los que acumulen la mayor cantidad de energía de la matriz original. (*) (5 pts)	b) Es conveniente realizar una reducción de dimensiones para lograr una mejor performance de ejecución de nuestro algoritmo de Machine Learning. (*) (5 pts)	c) SVD y PCA se enfocan en conservar la varianza, mientras que T-SNE se concentra en que los puntos cercanos sigan cercanos, y los puntos que originalmente estaban alejados se mantengan alejados. (*) (5 pts)	D) Cuando trabajamos con ISOMAP, la matriz de entrada de MDS tendrá siempre ceros en la diagonal. (*) (5 pts)
--	--	---	---

4) Luego de calcular 8 minhashes, se obtiene la siguiente tabla para 3 documentos distintos. Se pide encontrar los documentos candidatos si se usa el siguiente esquema: 2 AND, 2 OR, 2 AND

	D1	D2	D3
H1	1	1	1
H2	-1	1	1
H3	1	-1	1
H4	-1	-1	1
H5	-1	1	-1
H6	1	-1	-1
H7	1	1	1
H8	1	1	-1

(15 pts) (***)

5) Se tienen los siguientes documentos:

D1: VW VW FORD FIAT

D2: ALFA ROMEO LANCIA ALFA ROMEO

D3: FERRARI FIAT ALFA ROMEO LANCIA

D4: FORD FORD FIAT CHEVROLET FIAT

D5: CHEVROLET VW CHEVROLET VW VW

D6: FIAT FERRARI FERRARI

- Dada la consulta "FERRARI FIAT" dar el resultado de la consulta rankeadas utilizando TF.IDF.

- Considerando como relevantes los documentos que hablen únicamente sobre marcas italianas (Fiat, Ferrari, Alfa Romeo, Lancia), calcular la Precisión, Recall y F1 Score.

(****) (15 pts)

6) Discute un compresor nuevo para ser utilizado luego de aplicar MTF+BS a un archivo. Analiza y explícale en detalle el por qué de su propuesta. (****) (15 pts)

7) El Fiscal de Distrito Harvey Dent no está convencido de que la irrupción de Batman en Ciudad Gótica lo haya significado a la población y al departamento de policía una mejoría en la lucha contra el crimen organizado (categoría número 10 en el dataframe de casos).

Es tu misión ayudar al Comisionado Gordon planteando una visualización para demostrar a lo largo del tiempo como fue evolucionando la lucha contra el crimen partir de la participación de Batman, y el valor que le brinda al GCPD su ayuda. (****) (15 pts)

10 2018 - 10P

EJ (2)

CASOS POLICIALES

y-m-d
 (FECHA, ID-CASO, DESCRIPCION, ESTADO, CATEGORIA
 LONG, LAT) (1, 2, 3)

BATMAN

(ID-CASO, RESPUESTA)
 (0, 1)

(?)
 A) TASA DE RESOLUCIÓN DE CASOS DE LA POLICIA
 POR CATEGORÍA

CASOS POLICIALES y BATMAN = CASOS POLICIALES . MERGE (

BATMAN, HOW = LEFT, ON = ID-CASO)

FILTRO SIN BATMAN = CASOS POLY BATMAN [RTA] == 0

SOLO POLI = CASOS POLY BATMAN [FILTROSIN BATMAN]

RESULTADO = SOLO POLI . GROUP BY (['CATEGORIA'])
 [ESTADO_CASO] . AGG (LM X: (X == 2) . MEAN())

↓
 SOLO APLICAQ EL
 A ESTA COLUMNA

↓ CONVIERTO EN TRUE
 LOS CASOS CERRADOS
 Y EL RESTO EN FALSE

↓ SACO EL
 PROMEDIO

B) TASA DE RESOLUCIÓN DE CASOS CON AYUDA DE BATMAN

POLICONBATMAN = CASOS POLICIALES. MEERGE (BATMAN
HON = LEFT ON = ID_CASO)

FILTROCOBATMAN = POLICONBATMAN [ESTA] == 1

POLIFTBATMAN = POLICONBATMAN [FILTROCOBATMAN]

RESULTADO = POLIFTBATMAN [ESTADO - CASO]

• APPLY (LAMBDA X: X == 2) • MEAN()



LE APLICO A LA COLUMNA ESTADO

• LAMBDA X → EN ESTE CASO
CONVIERTETE A TRUE
Y FALSE

C) INDICAR EL MES DEL AÑO PASADO

USO EL DF POLIFTBATMAN

POLIFTBATMAN [FECHA] = PD. TO-DATETIME (ANO, MES, DIA)
POLIFTBATMAN [FECHA], FORMAT = 'Y.Y.Y.M/D'

FILTROAÑO PASADO = POLIFTBATMAN [FECHA].DT. YEAR == 2019

CASOSAÑOPASADO = POLIFTBATMAN [FILTROAÑO PASADO]

CASOSAÑOPASADO [FECHA].DT. MONTH . MODE ()

↓
EL
DE MODA

EJ (1)

VENTAS-VJ

(ID-VJ, ID-TIENDA, MÉS, AÑO, TOTAL-VENTAS-M)

o

1

2

3

4

TIENDAS

(ID-TIENDA, DIRECCION, LATITUD, LONG, CP, ESTADO)

o

1

2

3

4

5

A) TIENDA CON MENOS CANTIDAD DE VENTAS
EN EL ESTADO DE GEORGIA EN EL 2017

VENTASF = VENTAS-VJ.FILTER (LAMBDA X :
 $x[3] == 2017$)

DOQUE LAS VENTAS SON DOBLES EN LA TIENDA R

• VENTAS = VENTASF . MAP (LAMBDA X :
 $(x[1], x[4])$. REDUCE BY KEY (A+B)

TIENDASF = TIENDAS.FILTER (LAMBDA X :
 $x[5] == GEORGIA$)

- TIENDAS GEORGIA = TIENDASF . MAP (LAMBDA X :
 $(x[0], x[5])$)

AMBAS JUNTAS = TIENDAS GEORGIA . LEFT OUTER JOIN (VENTAS)

RESULTADO = AMBAS JUNTAS . REDUCE (LAMBDA A, B
A / IF A[1] < B[1] ELSE B).

EJ 5

DOC 1 = VW VW FORD FIAT

DOC 2 = ALFA ROMEO LANCIA ALFA ROMEO

DOC 3 = FERRARI FIAT ALFA ROMEO LANCIA

DOC 4 = FORD FORD FIAT CHEVROLET FIAT

DOC 5 = CHEVROLET VW CHEVROLET VW VW

DOC 6 = FIAT FERRARI FERRARI

A)

VW 1

CHEVROLET 5

VW

VW 1

VW 5

FORD FIAT

FORD 1

CHEVROLET 5

ALFA ROMEO

FIAT 1

VW 5

LANCIA

ALFA ROMEO 2

VW 5

CHEVROLET

LANCIA 2

FIAT 6

ALFA ROMEO 2

FERRARI 6

FERRARI 3

FERRARI 6

FIAT 3

ALFA ROMEO 3

LANCIA 3

FORD 4

FORD 4

FIAT 4

CHEVROLET 4

FIAT 4

	D1	D2	D3	D4	D5	D6	IDF
ALFA ROMEO	0	2	1	0	0	0	0,54
CHEVROLET	0	0	0	1	2	0	0,54
FERRARI	0	6	1	0	0	2	0,54
FIAT	1	0	1	2	0	1	0,24
FORD	1	0	0	2	0	0	0,54
LANCIA	0	1	1	0	0	0	0,54
VW	2	0	0	0	3	0	0,54

$$IDF = \log\left(\frac{N+1}{FT_i}\right) \quad N=6 = \log\left(\frac{7}{FT_i}\right)$$

Q = "FERRARI FIAT"

$$\bullet R(Q, D_1) = 1 * 0,24 = 0,24 \quad (1)$$

$$\bullet R(Q, D_2) = 0 \quad (\cancel{0})$$

$$\bullet R(Q, D_3) = 1 * 0,54 + 1 * 0,24 = 0,78 \quad (2)$$

$$\bullet R(Q, D_4) = 2 * 0,24 = 0,48 \quad (3)$$

$$\bullet R(Q, D_5) = 0 \quad (\cancel{0})$$

$$\bullet R(Q, D_6) = 2 * 0,54 + 1 * 0,24 = 1,32 \quad (1)$$

B) \hookrightarrow SI DA CERO, NO SE RECUPERA

DOCS RELEVANTES = DOC 2, DOC 3, DOC 6 = 3

DOCS NO RELEVANTES = DOC 1, DOC 4, DOC 5 = 3

RECUPERADOS = 4 DOC 1, DOC 3, DOC 4, DOC 6

NO RECUPERADOS = 2 DOC 2, DOC 5

RELEVANTES	RECUPERADOS	NO RECUP
	2 (A)	1 (B)
NO RELEVANTES	2 (C)	1 (D)

$$\text{PRECISIÓN} = \frac{2}{4} = \frac{1}{2}$$

$$\text{RECALL} = \frac{2}{3}$$

$$F_1 = \frac{2 \cdot PR}{P+R} = \frac{2 \cdot \frac{1}{2} \cdot \frac{2}{3}}{\frac{1}{2} + \frac{2}{3}} = \frac{4}{7}$$

EJ ④

1) 2 AND

AGREGO LA TABLA Y LA DIVIDO EN DOS. LOS CANDIDATOS DEBEN CUMPLIR EN AMBOS SUBCONJUNTOS

	D1	D2	D3		D2 y D3 CANDIDATOS
H1	1	1	1		H1 H2
H2	-1	1	1		D1 y D2 CANDIDATOS
H3	1	-1	1		H1 H3
H4	-1	-1	1		DEL OB NOS INTERESA QUE
H5	-1	1	-1		SOBRE NINGUNO UNO O OTRO
H6	1	-1	-1	(D2 D3) (D1 D2)	
H7	1	1	1		NO COINCIDE NINGUNO
H8	1	1	-1		NO HAY DOCS CANDIDATOS